TBSI 清华-伯克利深圳学院 Tsinghua-Berkeley Shenzhen Institute

Pick the Best Pre-trained Model: Towards Transferability Estimation for Medical Image Segmentation

Presenter: Jingyun Yang

2023/08/30

Background

- Sufficient annotated training samples are required for training while the labeling process of medical images is tedious and time-consuming.

- Transfer learning has been widely investigated to address the problem.

- Previous works mainly focused on the fine-tuning strategy to effectively adapt the knowledge from the pre-trained models to target tasks.
  - model repositories like Hugging Face and PyTorch Hub
  - these pre-trained models require less training time and have better performance and robustness

- Recent works observe that the pre-trained models cannot always benefit the downstream tasks.
  - when the knowledge is transferred from a less relevant source, it may not improve the performance or even negatively affect the intended outcome

Background

- Existing methods measure the task-relatedness between source and target datasets.
  - require source information available while medical images have more privacy and ethical issues and fewer datasets are publicly available than natural images.

- Directly measure the transferability of the pre-trained models without fully training based on the downstream/target dataset.
  - Log Expected Empirical Prediction (LEEP)
    - Utilized the log-likelihood between the target labels and the predictions from the source model.
  - Logarithm of Maximum Evidence (LogME)
    - Computed evidence based on the linear parameters assumption and efficiently leverages the compatibility between features and labels.
  - TransRate
    - Evaluated the transferability of models with the compactness and the completeness of embedding space.
  - Gaussian Bhattacharyya Coefficient (GBC)
    - Applied the Gaussian distribution to each class, and estimate the separability between classes as the basis for transferability estimation.

Challenges

- Previous works focused on classification and regression tasks without fully considering the properties of medical image segmentation.
  - C&R tasks can use a single n-dimensional feature vector to represent each image, segmentation problems lack a global semantic representation.
  - Propose class consistency to address the problem.

- Previous works focused on the relationship between the embeddings and downstream labels without exploring the effectiveness of the features themselves.
  - Propose feature variety to address the problem.

Challenges

- Medical images face severe class imbalance problems.
  - With excessive differences between foreground and background.
  - Only sample the foreground voxels with a pre-defined sampling number which is proportional to the voxel number of each class in the image.

- For semantic segmentation tasks, the feature pyramid is critical for the segmentation output of multi-scale objects while existing works neglect it.
  - Different decoders' outputs are used in the sliding window sampling process.

Framework

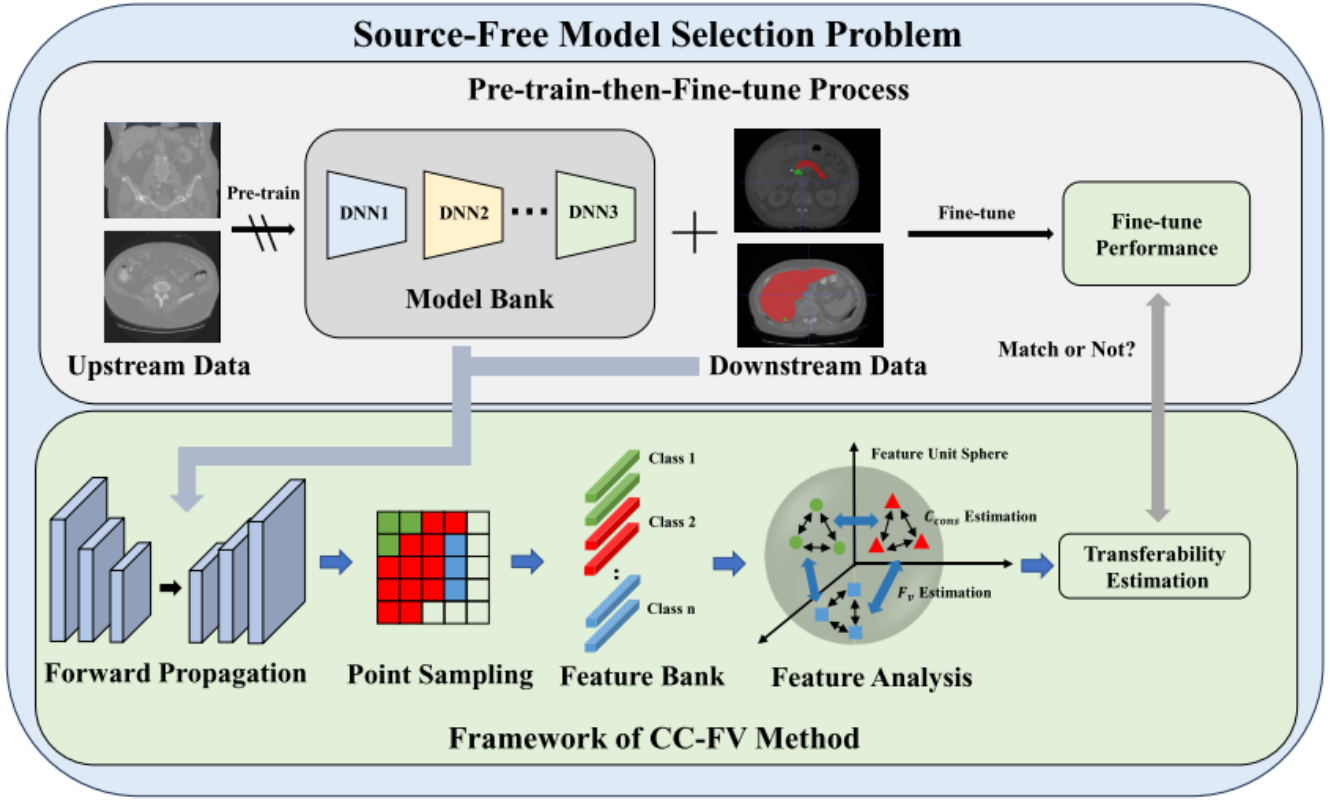


Fig. 1. Source-free model selection problem and the framework of our **C**lass **C**onsistency with **F**eature **V**ariety constraint(CC-FV) TE method. Our main goal is to predict the performance of models in the model bank after fine-tuning on downstream tasks without actually fine-tuning. Note that the upstream data are not available in our model selection process.

- Class consistency employs the distribution of features extracted from foreground voxels of the same category in each sample to model and calculate their distance.
- Feature diversity utilizes features sampled in the whole global feature map, and the uniformity of the feature distribution obtained by sampling is used to measure the effectiveness of the features themselves.

Class Consistency

- The pre-trained models are trained with specific pretext tasks based on the upstream dataset.
  - features extracted by the pre-trained models cannot perfectly distinguish the foreground and background of target data.

- If the features are generalizable, foreground region features will likely follow a similar distribution even without fine-tuning.
  - Features extracted by the pre-trained model should be consistent within the class of the target dataset.

Class Consistency

- Given a pair of target data X_j and $X_{j'}$:
  - The distribution of the features is modeled with the n-dimensional Gaussian distribution.
  - The class consistency between the data pair is measured by the Wasserstein distance:

$$\mathcal{W}_2^2(F_j^k, F_{j'}^k) = \left\|\boldsymbol{\mu}_{F_j^k} - \boldsymbol{\mu}_{F_{j'}^k}\right\|^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_{F_j^k}\right) + \mathrm{Tr}\left(\boldsymbol{\Sigma}_{F_{j'}^k}\right) - 2\,\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}_{F_j^k}\boldsymbol{\Sigma}_{F_{j'}^k}\right)^{1/2}\right) \quad (1)$$

where $\mu_{F_j^k}$, $\mu_{F_{j'}^k}$ are the mean of Gaussian distribution F_j^k, $F_{j'}^k$ and $\Sigma_{F_j^k}$ and $\Sigma_{F_{j'}^k}$ are covariance matrices of F_j^k and $F_{j'}^k$.

  - Calculate the wasserstein distance of the distribution with voxels of the same class in a sample pair comprised of every two samples in the dataset.
  - The pre-defined sampling number is proportional to the voxel number of each class in the image.

```
feature_dict = ms_sliding_window_sampling(layers, configs["sample_num"],
                                          val_data["data"], data_seg, [configs['roi_z'],
                                                                       configs['roi_y'], configs['roi_x']],
                                          configs['sw_batch_size'],
                                          model,
                                          overlap=configs['infer_overlap'],
```

```
"layers": ["decoder3", "decoder2", "out"],
"sample_num": {"decoder3": 100, "decoder2": 200, "out": 400},
"model": {
    "name": "UNETR"
},
"num_classes": 1,
"roi_x": 112,
"roi_y": 144,
"roi_z": 64,
```

- The class consistency is defined as:

$$C_{cons} = \frac{1}{N(N-1)} \sum_{k=1}^{C} \sum_{i \neq j} \mathcal{W}_2(F_i^k, F_j^k) \quad (2)$$

Feature Variety

- Class consistency is only concerned with local homogeneity of information while neglecting the integral feature quality assessment.
  - learn some trivial solutions
  - overfitted models have limited generalization capacity and are difficult to apply to new tasks

- Feature variety constraint measures the expressiveness of the features themselves and the uniformity of their probability distribution.
  - Highly complex features are not easily overfitted in the downstream tasks and do not collapse to cause a trivial solution.

Feature Variety

- To prevent overfitting and trivial features, we expect the distribution of features in the feature space to be as uniform and dispersed as possible.
- Employ the hyperspherical potential energy to measure the expressiveness of the features and the uniformity of their probability distribution.
  - $$E_s(v) = \sum_{i=1}^{L} \sum_{j=1, j\neq i}^{L} e_s\left(\left\|v_i - v_j\right\|\right) = \begin{cases} \sum_{i\neq j} \left\|v_i - v_j\right\|^{-s}, & s > 0 \\ \sum_{i\neq j} \log\left(\left\|v_i - v_j\right\|^{-1}\right), & s = 0 \end{cases} \quad (3)$$
    - v is sampled feature of each image with point-wise embedding v_i
    - L is the length of the feature, which is also the number of sampled voxels.
- For the dataset with N cases, the feature variety is formulated as:
  - $$F_v = \frac{1}{N} \sum_{i=1}^{N} E_s^{-1}(v) \quad (4)$$

Overall Estimation

- As for semantic segmentation problems, the feature pyramid structure is critical for segmentation results.
- The final transferability of pre-trained model m to dataset t is defined as:

$$\mathcal{T}_{m \to t} = \frac{1}{D} \sum_{i=1}^{D} \log \frac{\boldsymbol{F_v^i}}{\boldsymbol{C_{cons}^i}} \tag{5}$$

  - where D is the number of decoder layers used in the estimation.
  - decrease the sampling ratio in the decoder layer close to the bottleneck to avoid feature redundancy.

```
{
    "layers": ["decoder3", "decoder2", "out"],
    "sample_num": {"decoder3": 100, "decoder2": 200, "out": 400},
    "model": {
        "name": "UNETR"
    },
```

Experiment

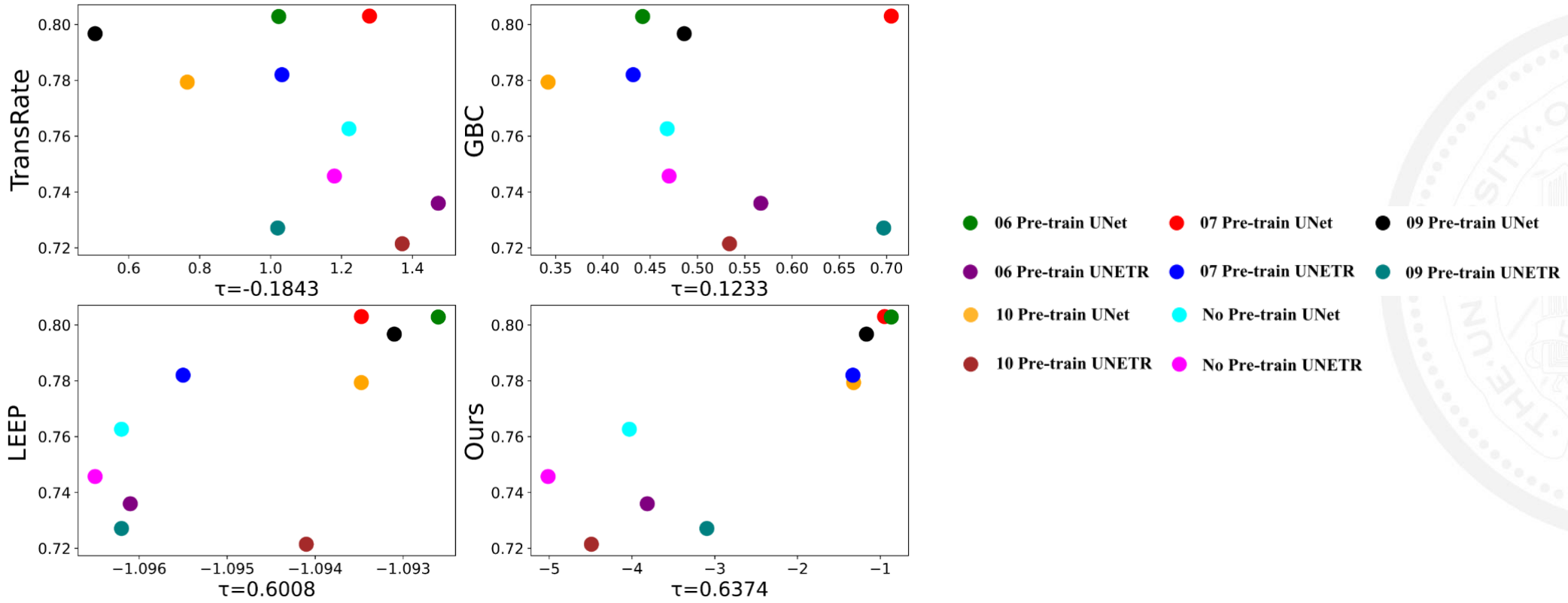
- Conduct experiments on 3D CT images of The Medical Segmentation Decathlon (MSD) dataset:
  - Task03 Liver: liver and tumor segmentation
  - Task06 Lung: lung nodule segmentation
  - Task07 Pancreas: pancreas and pancreas tumor segmentation
  - Task09 Spleen: spleen segmentation
  - Task10 Colon: colon cancer segmentation
- For each dataset:
  - use the other four datasets to pre-train the model
  - fine-tune the model on this dataset to evaluate the performance as well as the transferability
  - using the correlation between two ranking sequences of upstream pre-trained models.
- The baseline methods including TransRate, LogME, GBC and LEEP.

Metric

- Use weighted Kendall's τ and Pearson correlation coefficient for the correlation between the Transferbility Estimation (TE) results and fine-tuning performance.
- For Kendall's τ:
  - The Kendall's τ ranges from [-1, 1]
  - τ =1 means the rank of TE results and performance are perfectly correlated $\mathcal{T}_{s\to t}^{i} > \mathcal{T}_{s\to t}^{j}$ if and only if $\mathcal{P}_{s\to t}^{i} > \mathcal{P}_{s\to t}^{j}$
  - Assign a higher weight to the good models in the calculation, known as weighted Kendall's τ
- For Pearson coefficient:
  - The Pearson coefficient also ranges from [-1, 1].
  - Measures how well the data can be described by a linear equation.
  - The higher the Pearson coefficient, the higher the correlation between the variables.

Results

- Visualize the average Dice score and the estimation value on Task03 Liver/Tumor Segmentation.
  - The vertical axis represents the average Dice of the model
  - The horizontal axis represents the transferability metric results.
  - Standardize the various metrics uniformly



Results

Table 1. Pearson coefficient and weighted Kendall's τ for transferability estimation

Data/Method	Metrics	Task03	Task06	Task07	Task09	Task10	Avg
LogME	τ	-0.1628	-0.0988	0.3280	0.2778	-0.2348	0.0218
	pearson	0.0412	0.5713	0.3236	0.2725	-0.1674	0.2082
TransRate	τ	-0.1843	-0.1028	0.5923	0.4322	0.6069	0.2688
	pearson	-0.5178	-0.2804	0.7170	0.5573	0.7629	0.2478
LEEP	τ	0.6008	0.1658	0.2691	0.3516	0.5841	0.3943
	pearson	0.6765	-0.0073	0.7146	0.1633	0.4979	0.4090
GBC	τ	0.1233	-0.1569	0.6637	0.7611	0.6643	0.4111
	pearson	-0.2634	-0.3733	0.7948	0.7604	0.7404	0.3317
Ours CC-FV	τ	0.6374	0.0735	0.6569	0.5700	0.5550	**0.4986**
	pearson	0.8608	0.0903	0.9609	0.7491	0.8406	**0.7003**

- Most of the existing methods are not designed for segmentation tasks with a serious class imbalance problem.
- These methods rely only on single-layer features and do not make good use of the hierarchical structure of the model.

Ablation Study

- Analyze the impact of class consistency C_{cons} and feature variety F_v

Table 2. Ablation on the effectiveness of different parts in our methods

Data/Method	Task03	Task06	Task07	Task09	Task10	Avg
Ours CC-FV	0.6374	0.0735	0.6569	0.5700	0.5550	**0.4986**
Ours w/o C_{cons}	0.1871	-0.2210	-0.2810	-0.0289	-0.2710	-0.1230
Ours w/o F_v	0.6165	0.3235	0.6054	0.2761	0.5269	0.4697
Single-scale	0.4394	0.0252	0.5336	0.5759	0.6007	0.4341
KL-divergence	-0.5658	-0.0564	0.2319	0.4628	-0.0323	0.0080
Bha-distance	0.1808	0.0723	0.2295	0.7866	0.4650	0.3468

1. Though F_v can not contribute to the final Kendall's τ directly, C_{cons} with the constraint of F_v promotes the total estimation result.
2. Compare the performance of the method at single and multiple scales to prove the effectiveness of the multi-scale strategy.
3. KL-divergence and Bha-distance are unstable in high dimension matrics calculation and the performance is also inferior to the Wasserstein distance.

Ablation Study

- Visualize the distribution of different classes.

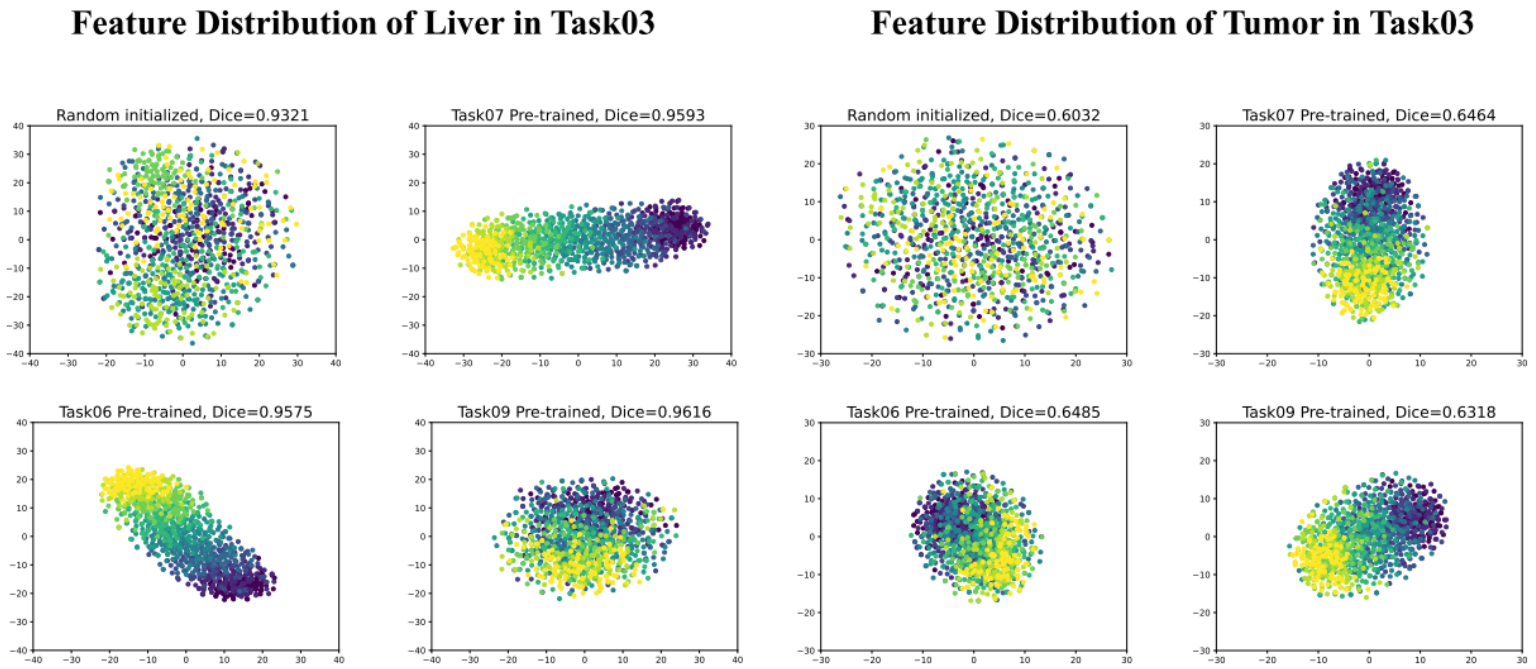


Fig. 3. Visualization of features with same labels using t-SNE. Points with different colors are from different samples. Pre-trained models tend to have a more consistent distribution within a class than the randomly initialized model and after fine-tuning they often have a better Dice performance than the randomly initialized models.

Contribution

- Propose a transferability estimation method based on class consistency with feature variety constraint
- Raise the problem of model selection for upstream and downstream transfer processes in the medical image segmentation task.
- Raise the problem of the ethical and privacy issues inherent in medical care and the computational load of 3D image segmentation tasks.

Reference

- Huang, L.K., Huang, J., Rong, Y., Yang, Q., Wei, Y.: Frustratingly easy transfer- ability estimation. In: ICML. pp. 9201–9225. PMLR (2022)
- Nguyen, C., Hassner, T., Seeger, M., Archambeau, C.: Leep: A new measure to evaluate transferability of learned representations. In: ICML. pp. 7294–7305. PMLR (2020)
- Pándy, M., Agostinelli, A., Uijlings, J., Ferrari, V., Mensink, T.: Transferability estimation using bhattacharyya class separability. In: CVPR. pp. 9172–9182 (2022)
- You, K., Liu, Y., Wang, J., Long, M.: Logme: Practical assessment of pre-trained models for transfer learning. In: ICML. pp. 12133–12143. PMLR (2021)

Thanks